ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414677

COUNT AND SEPARATE: INCORPORATING SPEAKER COUNTING FOR CONTINUOUS SPEAKER SEPARATION

Zhong-Qiu Wang[♪] *and DeLiang Wang*[♪,♫]

[♪]Department of Computer Science and Engineering, The Ohio State University, USA
[♫]Center for Cognitive and Brain Sciences, The Ohio State University, USA
wang.zhongqiu41@gmail.com, dwang@cse.ohio-state.edu

ABSTRACT

This study leverages frame-wise speaker counting to switch between speech enhancement and speaker separation for continuous speaker separation. The proposed approach counts the number of speakers at each frame. If there is no speaker overlap, a speech enhancement model is used to suppress noise and reverberation. Otherwise, a speaker separation model based on permutation invariant training is utilized to separate multiple speakers in noisy-reverberant conditions. We stitch the results from the enhancement and separation models based on their predictions in a small augmented window of frames surrounding an overlapped segment. Assuming a fixed array geometry between training and testing, we use multi-microphone complex spectral mapping for enhancement and separation, where deep neural networks are trained to predict the real and imaginary (RI) components of direct sound from stacked reverberant-noisy RI components of multiple microphones. Experimental results on the LibriCSS dataset demonstrate the effectiveness of our approach.

Index Terms—Complex spectral mapping, continuous speaker separation, microphone array processing, deep learning.

1. INTRODUCTION

Considerable progress has been made towards solving the talker-independent speaker separation problem, since deep clustering (DC) [1] and permutation invariant training (PIT) [2] were proposed to address the label permutation problem. To further improve separation, subsequent studies leverage microphone array processing [3]–[6], magnitude- and complex-domain phase estimation [7], [8], time-domain processing [9], and extra information such as speaker embeddings [10] and visual cues [11]. On wsj0-2mix and 3mix [1], a popular benchmark dataset containing monaural anechoic two- and three-speaker mixtures, current state-of-the-art approaches produce separation results that sound almost indistinguishable from clean speech, and the performance improvement measured by scale-invariant signal-to-distortion ratio is more than 20 dB over no processing [12].

The success on the wsj0-2mix and 3mix datasets however partly benefits from several strong assumptions. They may prevent the successful applications of many algorithms performing well on wsj0-2mix and similar datasets to realistic human conversations, where speaker overlap naturally happens. First, realistic recordings inevitably contain environmental noises and room reverberation. One should keep an eye on the robustness during algorithmic design. Second, the number of speakers is unknown beforehand and has to be estimated in order to do DC or PIT. In meeting scenarios, the number of speakers can vary from two to more than ten. Third, fully-overlapped speech as simulated in wsj0-2mix seldomly happens in natural conversations, and the overlap ratios among speakers can vary dramatically [4]. Most of the time, only one speaker talks. There could be short pauses or long silence between consecutive utterances. Sometimes another speaker interrupts. So, two-speaker overlap is common, but the case where more than two speakers talking at the same time rarely happens. Fourth, many studies assume offline-processing scenarios, where each utterance has been accurately segmented. In a streaming system, however, speech signals come as a continuous stream. How to modify the algorithms for online processing is an important problem to study.

In this context, a block-online approach is proposed in [13]–[15] to address continuous speaker separation, where speech signals from an unknown number of speakers, degraded by environmental noise, room reverberation and a wide range of speaker overlap, arrive as a continuous stream. These studies assume that in each fixed-length short processing block, typically 2.4-second long, there are at most two speakers talking, so that a two-speaker separation model based on for example utterance-wise PIT (uPIT) can be applied in each block for separation. Consecutive blocks are designed to be overlapped, and the overlapped regions is used for stitching the separation results in consecutive blocks [16].

Although working well most of the time when turn taking does not happen frequently [15], this overlapped-block approach makes a strong assumption that each processing block can only have two speakers at most. In addition, the length of the processing block cannot be long, because there could be more than two speakers in a longer block. However, using a small processing blocks limits the amount of contextual information a model can exploit. One can indeed train a say three- or four-speaker PIT model for block-online processing. However, it is unclear yet whether such a model would perform well when speakers are not fully overlapped and when the model is applied to process blocks with zero, one or two speakers.

Our study tackles continuous speaker separation from the angle of frame-wise speaker counting. Instead of assuming that there are at most two speakers in each short processing block, we assume that there are up to two speakers at each frame. We perform frame-wise three-class classification (i.e. zero, one or two speakers) for speaker counting. Our system includes a speech enhancement module as well as a speaker separation module. At run time, if the speaker counting module figures out that there is no speaker overlap, the enhancement module is picked for enhancement and if speaker overlap is happening, the speaker separation module is applied for separation. For one-speaker segments, although a multi-speaker model could put the speaker in one output and set the others to silence, we believe that a speech enhancement model should produce better results, as it only needs to enhance speech from noise and reverberation and does not have to learn to separate multiple speakers. We stitch the results from the enhancement and separation modules

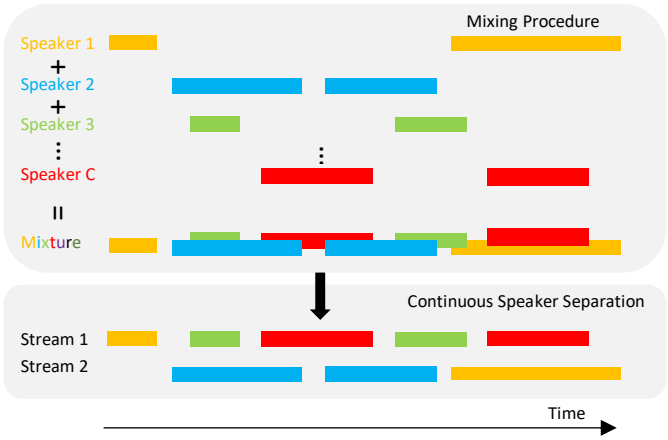

Figure 1. Task illustration.

based on a small augmented window surrounding the overlapped frames, where both modules need to make predictions in this common region. This way, we can potentially use a larger context window for our models and get rid of the two-speaker assumption in the block-online approaches. As an initial step towards online or low-latency continuous speaker separation, this study only deals with offline processing.

The rest of this paper presents the physical model and proposed algorithms in Sections 2 and 3, experimental setup and evaluation results in Sections 4 and 5, and conclusions in Section 6.

2. PHYSICAL MODELS AND OBJECTIVES

Given a P-channel conversational signal recorded in a noisy-reverberant environment with C speakers and assuming that there are at maximum two speakers at any time frame, our study aims at separating the mixture into two anechoic streams, each with no speaker overlap. Note that the total number of speakers C is assumed unknown and the signal could last minutes or tens of minutes. See Figure 1 for an illustration.

We assume a uniform circular array geometry, and that the same array is used in training and testing. In [15], it is shown that a separation model trained on a simulated array generalizes well to a real array with matched geometry.

3. PROPOSED ALGORITHMS

Our system contains three models, one for frame-wise speaker counting, one for speech enhancement, and one for speaker separation. If the speaker counting module finds that there is speaker overlap, the speaker separation module is applied for separation, and otherwise, the speech enhancement model is used to remove noise and reverberation. All of them use the same input features, and share an encoder-decoder network architecture (see Figure 3).

3.1.Speech Enhancement

Our enhancement network predicts the summation of the direct-path signals at the two streams, producing an estimate of the only speaker in one-speaker frames and the summation of the two speakers in overlapped frames. It is a multi-microphone input and single-microphone output (MISO) network [15] trained to predict the real and imaginary (RI) components of target speech at a reference microphone based on the mixture RI components at all the microphones and the mixture magnitude at the reference microphone. This

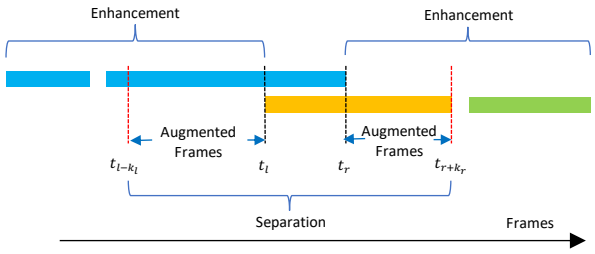

Figure 2. Illustration of frame stitching for CSS at run time.

network exhibits strong performance in tasks such as multi-microphone speech dereverberation [17] and speaker separation [15].

Following [18], [19], [15], the loss function is defined on the predicted RI components and their resulting magnitude

$$\mathcal{L}_{q,\text{Enh}} = \left\|\hat{R}_q - \text{Real}\left(\sum_{c=1}^{C} S_q(c)\right)\right\|_1 + \left\|\hat{I}_q - \text{Imag}\left(\sum_{c=1}^{C} S_q(c)\right)\right\|_1 + \left\|\sqrt{\hat{R}_q^{\,2} + \hat{I}_q^{\,2}} - \left|\sum_{c=1}^{C} S_q(c)\right|\right\|_1 \quad (1)$$

where $S_q(c)$ denotes the stream of the direct-path signal of speaker c at a reference microphone q, $\hat{R}_q$ and $\hat{I}_q$ are the predicted RI components produced by a linear activation in the output layer, $\text{Real}(\cdot)$ and $\text{Imag}(\cdot)$ respectively extract the real and imaginary components, $|\cdot|$ computes magnitude, and $\|\cdot\|_1$ computes the L_1 norm.

Note that this network is only trained to extract speech from noise and reverberation, and does not separate multiple speakers.

3.2. Speaker Separation

When speaker overlap is detected, we apply a MISO based network for speaker separation. The model is trained using uPIT [2] on two-talker mixtures. The loss function is

$$\mathcal{L}_{q,\text{uPIT}} = \min_{\psi_q \in \Psi} \sum_{c=1}^{2} \left(\left\|\hat{R}_q(\psi_q(c)) - \text{Real}\left(S_q(c)\right)\right\|_1 + \left\|\hat{I}_q\left(\psi_q(c)\right) - \text{Imag}\left(S_q(c)\right)\right\|_1 + \left\|\sqrt{\hat{R}_q(\psi_q(c))^2 + \hat{I}_q(\psi_q(c))^2} - |S_q(c)|\right\|_1 \right) \quad (2)$$

where Ψ denote the set of all the permutations of two sources and ψ_q is a permutation at a reference microphone q, and $\hat{R}_q$ and $\hat{I}_q$ are the predicted RI components.

3.3.Speaker Counting

As each frame is assumed to have at most two speakers, we perform three-class classification (i.e. zero, one and two speakers) for frame-wise speaker counting using a Softmax layer. The objective function is averaged cross-entropy loss weighted by the summation of mixture magnitude. On our own simulated two-talker reverberant mixtures, the accuracy of speaker counting is around 97%, which is sufficiently accurate.

3.4. Offline Continuous Speaker Separation

We then use the counting results to switch between speech enhancement and speaker separation for continuous speaker separation. Based on the frame-wise counting results, we can find segments

with zero, one or two speakers. For segments with zero or one speaker, we use the enhancement network output by putting the predicted speech at one stream and set the other to empty. For segments with two speakers, we first augment the detected overlapped segment with at most K (set to 100 in this study) frames on each side, then perform two-speaker separation on the augmented segment, and finally stitch the enhancement output with the separation output (i.e. sequential grouping) on each side based on their predicted magnitudes at the augmented frames. The augmented frames can also provide contextual information for the separation network to produce better separation, especially when the overlapped region is short. See Figure 2 for an illustration.

Note that we need to ensure that the augmented frames do not contain a third speaker. Suppose that the overlap is from frame t_l to t_r (see Figure 2), k_l is selected as the largest integer that is less than or equal to K and such that each of the frames from t_{l-k_l} to t_{l-1} only contains one detected speaker. The rationale is that if each frame in $[t_{l-k_l}, t_{l-1}]$ only has one speaker and $[t_l, t_r]$ contains speaker overlap, then $[t_{l-k_l}, t_{l-1}]$ should belong to one of the two overlapped speakers, as we assume that there are two speakers at most in each frame. k_r is selected in a similar way for the right side.

4. EXPERIMENTAL SETUP

We test our proposed algorithms on LibriCSS [13], a recently proposed dataset designed for continuous speaker separation. It has ten hours of conversational speech recorded by playing LibriSpeech signals through loud speakers in reverberant rooms. The task is to perform conversational speech recognition with room reverberation and various ratios of speaker overlap. There are ten one-hour sessions, each including six ten-minute mini-sessions with different speaker overlap ratios ranging from 0% to 40%, including 0S (no overlap with short inter-utterance silence between 0.1 and 0.5 seconds), 0L (no overlap with long inter-utterance silence between 2.9 and 3.0 seconds), and 10%, 20%, 30% and 40% overlaps. The recording device has seven microphones, with six of them uniformed placed on a circle with a 4.25 cm radius and one at the circle center. The distance from loud speakers to the array center ranges from 33 cm to 409 cm. There are two kinds of evaluations, utterance-wise and continuous-input evaluations. In the utterance-wise task, each utterance has been pre-segmented using ground-truth information. Frontend processing is expected to produce two streams. Both streams are scored and the lower word error rate (WER) is considered as the final WER. In the continuous-input task, each mini-session is segmented into 60- to 120-second long segments, each including 8 to 10 utterances from at most eight speakers. The task is to recognize all the utterances in each segment. Frontend processing for this task is expected to output a number of streams, each free of speaker overlap. The ASR backend scores all the streams and combines the decoding results to compute the final WER.

LibriCSS only contains testing data. We need to simulate our own training and validation data for separation. Our training data consists of 76,750 (~129 hours) seven-channel two-speaker mixtures with moderate room reverberation and weak air conditioning noise. Among all the frames, 12% have no speaker, 55% one speaker and 33% two speakers. The dry clean source signals are sampled from the *train-clean-{100,360}* set of LibriSpeech. Assuming the array geometry of the LibriCSS recording device, we use an RIR generator [20] to simulate seven-microphone RIRs. The reverberation time is sampled from the range [0.2,0.6] s. The average distance between speaker and array center is sampled from the range [0.75,2.5] m. The two speakers are constrained to be at least 10° apart and their relative

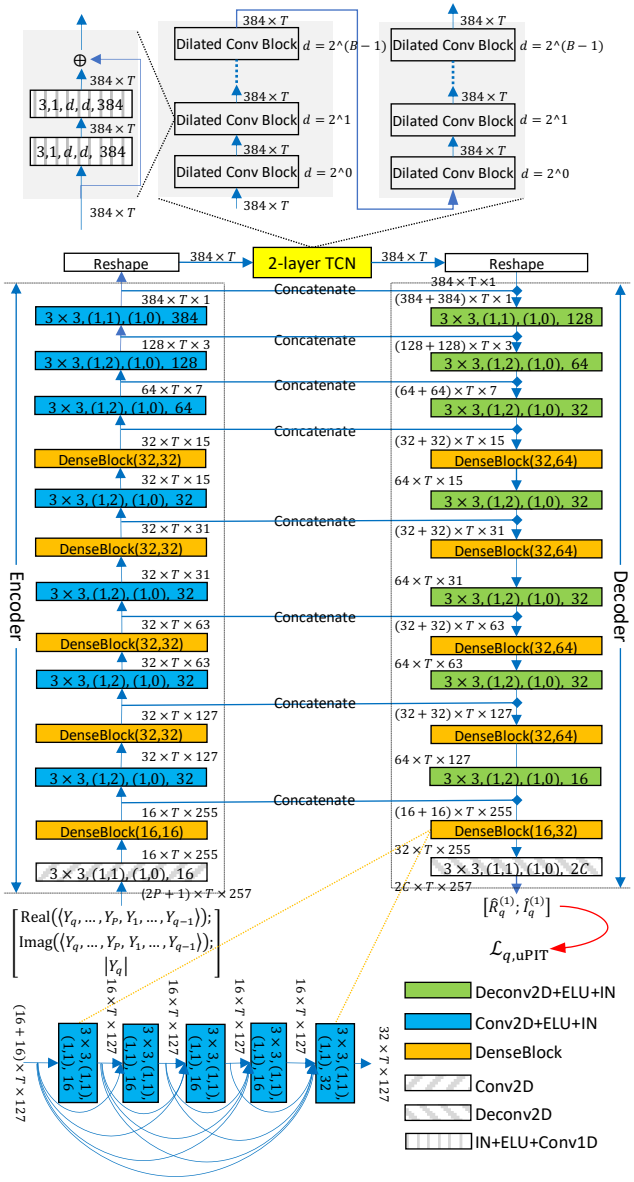

Figure 3. Network architecture of MISO for predicting RI components of S_q from multi-channel inputs $\boldsymbol{Y}$ and $|Y_q|$. The tensor shape after each encoder-decoder block is in the format: *featureMaps×timeSteps×frequencyChannels*. Each one of Conv2D, Deconv2D, Conv2D+ELU+IN and Deconv2D+ELU+IN blocks is specified in the format: *kernelSizeTime×kernelSizeFreq, (stridesTime, stridesFreq), (paddingsTime, paddingsFreq), featureMaps*. Each DenseBlock (g_1, g_2) contains five Conv2D+ELU+IN blocks with growth rate g_1 for the first four layers and g_2 for the last layer. The tensor shape after each TCN block is in the format: *featureMaps×timeSteps*. Each IN+ELU+Conv1D block is specified in the format: *kernelSizeTime, stridesTime, paddingsTime, dilationTime, featureMaps.*

energy level is sampled from [−5,5] dB. For each reverberant two-talker mixture, we draw an air conditioning noise from the REVERB corpus [21]. The SNR between each anechoic two-speaker mixture and noise is sampled from [5,25] dB. The labels used for training the speaker counting model is obtained by first applying a pre-trained

Table I
WER (%) on LibriCSS (Continuous-Input Evaluation, 7ch).

Approaches	Type	Overlap Ratio (%)					
		0S	0L	10	20	30	40
Unprocessed	-	15.4	11.5	21.7	27.0	34.3	40.5
Two-Speaker Separation	Block-Online	10.0	12.1	10.1	11.9	14.0	15.6
Proposed	Offline	**8.0**	**8.5**	**8.7**	**10.1**	**12.3**	**14.7**
Chen *et al.* [13]	Block-Online	11.9	9.7	13.4	15.1	19.7	22.0
Chen *et al*. [14]	Block-Online	11.0	8.7	12.6	13.5	17.6	19.6

Table II
WER (%) on LibriCSS (Continuous-Input Evaluation, 1ch).

Approaches	Type	Overlap Ratio (%)					
		0S	0L	10	20	30	40
Unprocessed	-	15.4	11.5	21.7	27.0	34.3	40.5
Two-Speaker Separation	Block-Online	12.5	17.4	13.4	16.4	20.8	23.7
Proposed	Offline	**9.2**	**8.2**	**11.5**	**15.1**	**19.2**	**22.4**
Chen *et al.* [13]	Block-Online	17.6	16.3	20.9	26.1	32.6	36.1
Chen *et al*. [14]	Block-Online	13.3	11.7	16.3	20.7	25.6	29.3

Table III
WER (%) on LibriCSS (Utterance-Wise Evaluation, 7ch).

Approaches	Type	Overlap Ratio (%)					
		0S	0L	10	20	30	40
Unprocessed	-	11.8	11.7	18.8	27.2	35.6	43.3
Two-Speaker Separation	Block-Online	6.8	7.1	7.9	**10.2**	**11.6**	**13.8**
Proposed	Offline	**5.7**	**5.9**	**7.6**	**10.2**	12.5	16.1
Chen *et al.* [13]	Offline	8.3	8.4	11.6	16.0	18.4	21.6
Chen *et al*. [14]	Offline	7.2	7.5	9.6	11.3	13.7	15.1
Oracle direct sound	-	4.9	5.1	-	-	-	-

Table IV
WER (%) on LibriCSS (Utterance-Wise Evaluation, 1ch).

Approaches	Type	Overlap Ratio (%)					
		0S	0L	10	20	30	40
Unprocessed	-	11.8	11.7	18.8	27.2	35.6	43.3
Two-Speaker Separation	Block-Online	9.5	8.9	11.9	15.9	**20.6**	**23.9**
Proposed	Offline	**7.4**	**6.8**	**10.6**	**15.8**	20.9	26.2
Chen *et al.* [13]	Offline	12.7	12.1	17.6	23.2	30.5	35.6
Chen *et al*. [14]	Offline	12.9	12.2	15.1	20.1	24.3	27.6

DNN based voice activity detection (VAD) model [22] to the anechoic signal of each stream and then combine the two VAD results.

Figure 3 illustrates the network architecture of our enhancement and separation models. It is a temporal convolutional network (TCN) sandwiched by a U-Net and includes an encoder for down-sampling and a decoder for up-sampling along frequency. DenseNet blocks are inserted at multiple frequency scales in the encoder and decoder. The rationale of this network design is that U-Net can maintain local fine-grained structure via its skip connections and exploit contextual information along frequency through down- and up-sampling, TCN can model long-range information via its dilated convolutions along time, and DenseNet blocks encourage feature reuse and improve discriminability. We use exponential linear units (ELU) in the activation layers and instance normalization (IN) in the normalization layers. Similar architectures exhibit strong performance in a number of tasks including speaker separation [8], [15], speech dereverberation [18] and speech enhancement [19]. The RI components of multiple microphones are stacked as features maps in the network input and output. We also include the mixture magnitude at the reference microphone as the input features, as it leads to more robust speaker separation and counting in realistic conditions [15]. Each network contains around 6.9 million parameters. The network atchitecture of speaker counting is similar to Figure 3, but without the decoder. The Softmax layer is added on top of TCN.

For offline processing, we normalize the sample variance of each multi-channel recording to unit variance before any processing. This normalization can deal with random gains in inputs, and is reported to be important for mapping-based approaches [18], [19]. The frame length is 32 ms and frame shift is 8 ms. The square root of Hann window is used as the analysis window. The sampling rate is 16 kHz. A 512-point discrete Fourier transform is applied to extract 257-dimensional complex spectra. Global mean-variance normalization is applied to all the input features.

Our study focuses on separation. We use the default ASR backend provided in LibriCSS for recognition to facilitate comparison with or by other studies. We perform signal resynthesis before extracting features for recognition.

5. EVALUATION RESULTS

Table I and Table II respectively report WER on the seven- and one-microphone continuous-input task of LibriCSS. As a block-online baseline, we use our two-speaker separation network trained based on multi-microphone complex spectral mapping for block-online processing, where the block size is 2.424 second, the block shift is 1.2 second, and the overlapped frames between consecutive blocks are used for block stitching. Our block-online system shows clearly better performance over the block-online systems in [13], [14]. For example, 15.6% vs. 22.0% and 19.6% in the 40% overlap condition in the seven-microphone case. However, in one-speaker conditions such as 0S and 0L, we find that this system cannot assign the target speaker to one stream and set the other to silence, i.e. there is some energy leakage from the higher-energy stream to the weaker one, resulting in degradation in WER performance. By using a speaker counting module to switch between enhancement and separation, and therefore avoiding producing two separation outputs in one-speaker segments, the proposed approach obtains clear improvement, especially in the 0S and 0L conditions.

Table III and Table IV report the WER on the seven- and one-microphone utterance-wise task of LibriCSS. Compared with the two-speaker model applied in a block-online manner, the proposed approach obtains clear improvements in the 0S and 0L conditions, which align with the findings in Table I and Table II, and comparable results on the overlap conditions. Both of them show overall better performance than [13] and [14].

6. CONCLUSION

We have proposed to use a speaker counting module to switch between speech enhancement and speaker separation for offline continuous speech separation. Evaluation results on the LibriCSS dataset indicate the effectiveness of our proposed approach. Future research shall deal with the case where there can be more than two speakers at any given frames and modify the proposed algorithms for online real-time processing.

7. ACKNOWLEDGMENT

This research was supported by an NSF grant (ECCS-1808932) and the Ohio Supercomputer Center.

8. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-Talker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1–5.

[4] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-Microphone Neural Speech Separation for Far-Field Multi-Talker Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5739–5743.

[5] Z.-Q. Wang and D. L. Wang, "Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[6] R. Gu *et al.*, "Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7319–7323.

[7] Z.-Q. Wang, K. Tan, and D. Wang, "Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 71–75.

[8] Y. Liu and D. L. Wang, "Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] Q. Wang *et al.*, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proceedings of Interspeech*, 2019, vol. 2019-Septe, pp. 2728–2732.

[11] A. Ephrat *et al.*, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Trans. Graph.*, vol. 37, no. 4, Apr. 2018.

[12] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End Speech Separation by Speaker Clustering," in *arXiv preprint arXiv:2002.08933*, 2020.

[13] Z. Chen *et al.*, "Continuous Speech Separation: Dataset and Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7284–7288.

[14] S. Chen *et al.*, "Continuous Speech Separation with Conformer," in *arXiv preprint arXiv:2008.05773v1*, 2020.

[15] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-Microphone Complex Spectral Mapping for Utterance-Wise and Continuous Speaker Separation," *arXiv Prepr. arXiv2010.01703*, 2020.

[16] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 241–245.

[17] Z.-Q. Wang and D. L. Wang, "Multi-Microphone Complex Spectral Mapping for Speech Dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 486–490.

[18] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[19] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[20] E. Habets, "Room Impulse Response Generator," 2010.

[21] K. Kinoshita *et al.*, "A Summary of the REVERB Challenge: State-of-The-Art and Remaining Challenges in Reverberant Speech Processing Research," *EURASIP J. Adv. Signal Process.*, no. 1, pp. 1–19, 2016.

[22] V. Manohar, "https://kaldi-asr.org/models/m4," 2018.